LLNL-PROC-659520

Bacterial Genome Reference Databases: Progress and Challenges

T. Slezak

September 2, 2014

November 2013 PDA Meeting
Bethesda, MD, United States
November 12, 2013 through November 15, 2013

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

LLNL-PROC-659520

November 2013 PDA Meeting Summary Article

Bacterial Genome Reference Databases: Progress and Challenges

Technical Point of Contact Information
Name: Tom Slezak
Email: slezak1@llnl.gov
Phone: (925) 422-5746
Address: 7000 East Avenue
Mail Code 174
Livermore, CA 94550

Tom Slezak Nov 2013 PDA Meeting summary article

Bacterial Genome Reference Databases: Progress and Challenges

Abstract

Accurate and sensitive detection of microbes against a complex background is a problem common to multiple aspects of human health, such as vaccines and other biologicals safety, blood safety, and diagnosing infectious diseases in humans or other hosts. The microbes in question could be bacterial, viral, fungal, or parasitical. To defend against such a broad array of microbes of potential safety concern, we need more than single-target PCR assays. Technologies such as highly-multiplexed PCR, broad-spectrum DNA/RNA microarrays, and next generation sequencing (NGS) are all potentially capable to provide increased protection against microbial contamination. Regulatory processes are currently struggling to keep up with rapid advances in all of these technologies, each of which is firmly based upon nucleic acid sequencing resulting in generation of megabases of data. A major question is the level of quality required for genomic data and metadata for the reference databases that are needed to allow these technologies to be developed, validated, and then used for front-line protection of human health. The background of this general problem is discussed and one example of an ongoing effort to establish quality levels for a bacterial genome reference database is presented.

Keywords

Genome reference databases, metadata, multiplex-PCR, broad-spectrum microarrays, next generation sequencing

Introduction

During the course of the November, 2013 PDA meeting we established the need for broad-spectrum technologies and sensitive assays to detect adventitious viruses in biological products. Several speakers also noted that bacterial and fungal contaminants are also of potential concern (although not the focus of this meeting.). Parallel efforts to determine bacterial agents in human clinical samples are underway and will be descrbed here.

Nuclei acid (NA)-based detection techniques are currently the most viable candidates for practical use for large-scale screening and confirmation of adventitious microbial agents of any type. These include:

- Single-plex PCR assays: Rapid and sensitive; based on specific pre-determined target sequences in known agents
- Multiplexed PCR and array-like devices: Can detect multiple known agents in a single test; based on specific pre-determined target regions
- NGS (NextGen Sequencing): Can detect both known and unknown agents without prior target sequence knowledge; requires extensive post-sequencing bioinformatics to determine agent(s) present.

Protein-based analyses, ranging from lateral-flow detection devices to mass-spectrometry, may be viable for some aspects of adventitious agent detection, but for lack of sensitivity, completeness of coverage, and ability to detect novel pathogens they are not total substitutes for NA-based techniques.

Dependence on Genomic Reference Databases

All NA-based detection techniques are dependent upon high-quality genomic reference databases. PCR and microarrays depend on sophisticated front-end bioinformatics to compare all available target and non-target genomes from reference databases to identify and select specific genomic regions that are highly conserved in target genomes and absent in non-target organisms. Frequently, strain variation will preclude the determination of sufficient regions that are totally conserved in *all* target strains; in such cases, sets of signatures may be needed, such that all target strains can be detected by the total set of PCR signatures and/or microarray probes. NGS uses genomic reference databases in a back-end fashion, comparing raw sequence reads or assembled “contigs” against the reference genomes to determine the most probable organism(s) present in the sample.

Historically, single-plex PCR assay validation was totally a lab-based process. The assay being validated would be subjected to many lab tests against a target-strain panel, a near-neighbor panel, and an “environmental” or “zoo” panel. As a more concrete example, validating a new *Bacillus anthracis* assay would involve testing against a panel of many dozens of BA strains, plus a large set of strains from other *Bacillus* species (*cereus*, etc.), plus a set of “other” organisms (that might include human, mouse, fly, various non-*Bacillus* organisms, soil and aerosol samples, etc.)

Genomic databases will be required for (eventual) FDA device validation

The accepted model of single-plex PCR assay validation simply does not scale for more information-rich detection technologies, such as highly-multiplexed PCR, DNA microarrays, or NextGen Sequencing (NGS). Consider even a modest 20-plex PCR assay that has one assay for each of 20 organisms. Each assay must not only be tested individually as per normal single-plex validation, but in the limit case you would also need to demonstrate that all possible combinations of multiple organism detection work properly. For a DNA detection microarray that has probes to identify some ~5,000 organisms the exhaustive lab validation combinatorics are truly astronomical. Finally, given that NGS could be given any possible DNA or RNA mix as input, it is clear that some combination of *in silico* and lab testing to achieve validation will be required for practical applications of the technologies providing broad microbial agents assays for improving human health care.

We will focus here primarily on bacterial genome reference databases. However, much of the discussion will also apply to viral, fungal, or parasitical reference databases.

What does “high-quality genomic bacterial reference database” mean in this context?

One operational definition might be: *Sufficient coverage of the bacterial strains circulating in nature such that species/strains pathogenic to humans can be clearly differentiated from non-pathogenic ones based on DNA sequence interrogation.* The end-goal is to provide definitive species and probable closest-sequenced-strain calls for unknown bacterial samples that could be either from isolates or complex samples (e.g., assembled contigs or mapping of raw reads).

Pathogen detection via multiplexed PCR, DNA microarrays, and NGS are all based on interrogating regions of pathogen genomes that are known or thought to be indicative of taxonomic resolution and/or factors of virulence or resistance. PCR and microarrays perform their bioinformatics up front so that the most informative regions can be properly targeted. NGS performs its bioinformatics after the fact, sifting through millions to billions of sequence reads to look for the differentiating regions. Whether the critical bioinformatics are performed pre- or post-experiment, it is clear that the quality of the reference database being used is of paramount importance. The complex software needed to perform comparative genomic analyses for multiplexed PCR, arrays, and NGS are beyond the scope of this article, but it is worth noting that the recent explosion of genomic sequence data has caused there to be many scaling issues.

Sufficient coverage has not generally been achieved for most pathogens of interest, including bacteria. There are also reasons to be concerned about the accuracy of sequencing and assembly of the bacterial strains in the database (http://genomebiology.com/content/9/3/R55).

Taxonomic fluidity must be dealt with in genomic reference databases. Perhaps the best example is for viruses, where the International Committee on Taxonomy of Viruses (ICTV) can take years to classify novel viruses, or correct mis-classification. ICTV has a large backlog of many thousands of "unclassified viruses" that practical systems must somehow deal with. Bacteria have also been reclassified/renamed frequently. (For two examples see: http://www.nphl.org/documents/fall2009whatsinaname.pdf and http://www.rhizobia.co.nz/taxonomy/rhizobia)

Ideally, a genomic reference database should be able to track taxonomic changes: what was the version of reference DB and taxonomy on the date when a particular analysis was performed? In practice, this is non-trivial to implement without opening the door for potential error creation. For example, you must remember the obsolete taxonomic designation to handle historical questions, but you must avoid returning information from the obsolete name for most (but not necessarily all) current queries.

Database curation is in the eye of the beholder

As exciting as it is to deal with the genomic data, the associated metadata is even more of a challenge. Both viewpoints and needs differ greatly among the various potential end-users of genomic data. Consider for a moment the viewpoints and needs of {assay developers; regulators; product manufacturers; doctors; law enforcement/military; organism researchers; ecologists; etc.} Each of these user categories has metadata that they are passionate about requiring, but the overlap between the categories can be rather depressingly small. Even where there is apparent overlap, the resolution scale or metrics may be completely different. As an example: What does "location" mean with respect to the genome of a bacterial organism? We have seen GPS coordinates, latitude/longitude, zip code, county, state, and country to name just a few. Sometimes the "location" is where the sample originated, other times it is the lab where the sample was processed. The "Ames strain" of *Bacillus anthracis* is perhaps the best known such example: the sample actually originated from a dead cow in southwest Texas (another “location” resolution example) but was eventually distributed widely from a lab in Ames, Iowa. Did the sample come from a host who died or one who recovered from a relatively mild illness? All too often, this type of metadata is not available.

Dates are another metadata nightmare. Consider what dates may apply to one sample whose genome is in the reference database (this is by no means a complete list):

- The date the sample was taken (from a host or from the environment)
- The date the sample was acquired (may be from a chain of “owners/distributors”)
    - Potential passage numbers and date(s); in media or in a host?
- The date the sample was processed (e.g., nucleic acid extraction; may be days to years after acquisition)
- The date the sample’s DNA was sequenced (may be long after extraction)
- The date the sample’s genome was “completed” (may be long after sequencing)
- The date the genome was submitted to the database (may be long after completion)
- The date of a related publication (often is tied to submission but not always)
- The last date that the genome or associated metadata was updated

You can easily see how some of these metadata may be vitally important to some users and of confusing irrelevance to others. It is seemingly inevitable that each user community involved in setting up a reference database will evolve their own definition of required and optional metadata. A formal attempt at establishing metadata standards exists (http://gensc.org/projects/mixs-gsc-project/ and http://wiki.gensc.org/index.php?title=MIGS/MIMS) but in actual practice one has to take what is available. This can lead to enormous metadata input forms that then tend to get filled out with almost no information since virtually all fields end up being “optional”. (Note that in some contexts, the “location” of origin of a particular sample may be sensitive information that will not be stored in the reference database, even though the genome itself may be allowed to be in the public domain.) One cannot retrospectively “fill in” missing metadata in any meaningful way. Sadly, it is also extremely difficult to query metadata meaningfully in genome reference databases due to both the sparseness of entries and the inconsistencies of what has been entered. (How many different spellings exist of your favorite institution or researcher in your favorite public reference genome database?)

Religious wars have also been fought over the “annotation” of genomic data. Genes, SNPs, VNTRs, rearrangements, insertion elements, transposons, and many other genomic features can be “annotated” onto the genomic sequence. Today, most of these features will come from computational predictions while some may actually be validated via lab experiments. Note that annotation of low-level features by itself may be insufficient for some users of genomic reference databases: “Is this a fully-virulent strain of *Bacillus cereus*?” may be difficult to determine from an annotation parts-list alone.

A likely outcome for genome reference databases?

The standards for data and metadata for both bacterial and viral reference databases to be used for vaccine safety, blood safety, and human medical diagnostics will likely have some differences around common cores. Differing needs for issues such as data quality versus completeness of strain coverage may cause differences in the genomes used, and different requirements for metadata will similarly cause regrettable but necessary variation.

A problem with using consensus sequences?

For viral reference genome databases one can ask whether viral consensus sequences are really sufficient or if quasi-species will need to be accommodated as the lowering cost of deep sequencing makes it possible to quantify the rare variants within a viral population.

Similarly, we are now realizing that bacteria also exist in quasispecies (http://www.pnas.org/content/102/27/9535.long
and http://www.ncbi.nlm.nih.gov/pubmed/10066468) Mapping to a single "best" reference genome database strain may be an oversimplification that we have to eventually deal with at a higher quasispecies resolution, particularly with respect to microbiomes.

Why aren't existing Genome Reference Databases sufficient?

Existing bacterial and viral genomic databases are insufficient for use as reference databases. The NCBI RefSeq genomes are of extremely high quality but their strain coverage is insufficiently broad. The economics of current NGS approaches do not lower the high total cost of finishing genomes at the absolute highest possible quality level; therefore, most genomic sequencing of bacteria is now draft that may never be taken to a finished state.

The NCBI nt (nucleotide) and WGS (whole genome sequence) databases contain draft genomes and thus have broader strain coverage. However, serious issues exist about the quality of both the sequence data and metadata. For one thing, these databases contain both sequencing and assembly quality differences stemming from 20+ years of evolution. This leads to inconsistencies and errors. (http://bioinformatics.oxfordjournals.org/content/21/24/4320.full)

What efforts are underway to develop new Genome Reference Databases?

There currently are multiple efforts in different parts of FDA that address the need for improved genomic reference databases:

- A DTRA-funded effort working with Peyton Hobson at FDA/CDRH with active involvement from NCBI, CDC, NMRC, NIST and others. This effort's focus is to determine the quality of genomic databases required to further the use of genomic sequencing as a human diagnostic technique. This effort includes an ongoing pilot project centered upon resolving different bacteria strains within a human DNA background. The database(s) that are developed using the results of this project will support the future use of validated broad-spectrum and sensitive assays for the detection of microbial agents in human clinical samples. This could include multiplexed PCR, DNA microarrays, and NGS.

- A blood safety effort, led by Sanjai Kumar, FDA/CBER and including the blood supply industry is focusing on viruses. They held a workshop in April, 2013: Application of Advances in Nucleic Acid and Protein Based Detection Methods for Multiplex Detection of Transfusion-Transmissible Agents and Blood Cell Antigens in Blood Donations. This meeting stressed the need for a high-quality viral reference database relevant to blood safety needs.
- An effort focusing on vaccines and biologicals safety, led by Arifa Khan, FDA/CBER and the Parental Drug Association (PDA) is focusing on both pathogenic and nonpathogenic viruses, which could be potential contaminants in biologicals,as noted within these proceedings. These efforts include active national and international participation from industry (biological and biotechnological), academia, and various government agencies including NCBI, NIST, and NIBSC, and contract research organizations.

NCBI has graciously volunteered to act as a repository of high-quality bacterial and viral reference databases for these kinds of projects, and are actively participating. As the reference databases come into being, many questions will need to be resolved. Will the databases be mirrored internationally? If separate reference database versions are required for different communities, how will these be handled? How will the databases deal with versioning? (e.g., users should be able to request to download the exact set of genomes that were available as of a specific point in time, so that results can be compared equally.)

Reference Genomic Database Standards Must Be Practical

Many sequencing centers today list the cost of NCBI submission as being the most expensive part of the sequencing process due to the drastic reduction in sequencing cost (personal communication). We cannot require additional metadata fields that will not actually be regularly used without risking that increasing numbers of genomes either will not be submitted at all or else will be greatly delayed being released.

Efforts to standardize metadata exist, but may be focused on particular end-uses of the genomic data. (http://www.ncbi.nlm.nih.gov/pubmed/19850722
and http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1001088) For example, the NIAID-funded PATRIC (PathoSystems Resource Integration Center) database has defined 61 metadata attributes requested for all samples. This is an example of an *ad hoc* standard that appears to be well thought-out but specific to one database. (http://enews.patricbrc.org/faqs/genome-metadata-faqs/)

You can't get retrospective metadata

For existing genomes in public databases, vital details of the sequencing platform, chemistry used, and the assembly pipeline used may not be available; much less the version numbers of multiple tools, hundreds of underlying libraries, operating systems, etc. that comprise the complex (and always evolving) software suite inherent with genomic sequencing.

Similarly, information about the origin of the strain, including the host it came from and the number of subsequent passages prior to sequencing, may not be available. Note that many bacteria undergo large genomic changes (e.g., loss of plasmids and genes that may encode virulence factors) when propagated in artificial laboratory conditions.

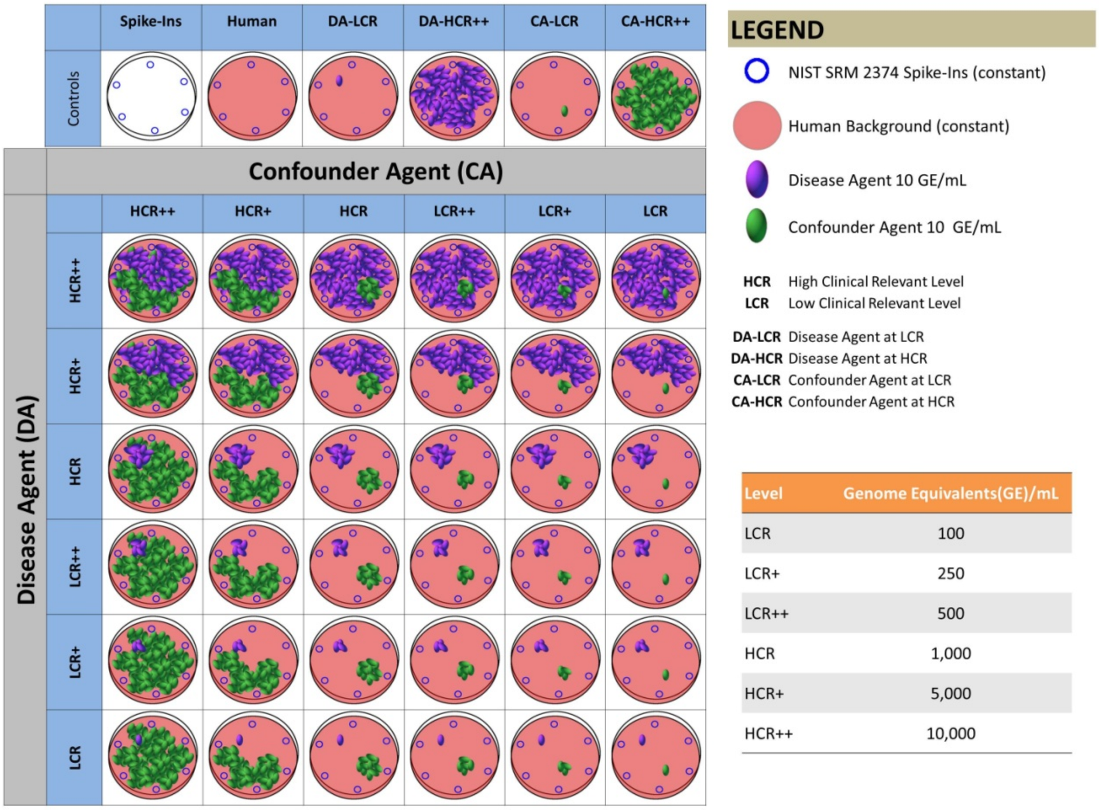
We must therefore accept existing genomic data and available metadata as usable (perhaps with some level of QC checking) or else we are faced with needing to re-sequence *everything* from scratch. An unfortunate corollary of this inability to get retrospective metadata is that queries based on metadata are thus inherently incapable of providing optimal answers unless they are limited to metadata fields that are present for all genomes covered by the query.

Experience of a Pilot Bacterial Genome Reference Database Quality Project

As mentioned above, DTRA is funding a pilot bacterial reference database project with the FDA and others involved (NMRC, NIST, CDC, UTMB, etc.) This Bacterial pilot is looking at mock human clinical samples spiked with pathogenic agent (*S. aureus*) and confounder (*S. epidermis*) at multiple levels. The study is aimed at answering the following two questions:

- Can confident species discrimination be performed by current NGS at levels of pathogen presence approaching low clinical-relevance?
- What is the right quality of reference database needed to utilize NGS as a clinical diagnostic tool?

The diagram below is the sample matrix for the pilot project:


Spike-Ins
Human
DA-LCR
DA-HCR++
CA-LCR
CA-HCR++
Controls
Confounder Agent (CA)
HCR++
HCR+
HCR
LCR++
LCR+
LCR
Disease Agent (DA)
HCR++
HCR+
HCR
LCR++
LCR+
LCR
LEGEND
NIST SRM 2374 Spike-Ins (constant)
Human Background (constant)
Disease Agent 10 GE/mL
Confounder Agent 10 GE/mL
HCR High Clinical Relevant Level
LCR Low Clinical Relevant Level
DA-LCR Disease Agent at LCR
DA-HCR Disease Agent at HCR
CA-LCR Confounder Agent at LCR
CA-HCR Confounder Agent at HCR

Level	Genome Equivalents(GE)/mL
LCR	100
LCR+	250
LCR++	500
HCR	1,000
HCR+	5,000
HCR++	10,000

Figure 1. DTRA/FDA Bacterial Reference Database pilot study sample matrix (Diagram courtesy of Heike Sichtig, FDA)

Four levels of genomic reference databases will be utilized for making calls of what is present:

Database 1: The genomes only for the 2 spiked-in strains used in the study

Database 2: Database 1 plus all other finished genomes of *Staph aureus* and *Staph epidermis*

Database 3: Database 2 plus all draft genomes of *Staph aureus* and *Staph epidermis*

Database 4: All available finished and draft bacterial genomes

The logic behind these different databases may require some explanation. Database (1) represents the absolute “best” case for attempted species determination of the 2 strains used in the study. If this database cannot resolve whether one or both of the species are present using their exact genomes, then no other database can do better. Databases (2) and (3) introduce all other finished and draft genomes of *S.aureus* and *S. epidermis*, allowing us to determine if these additional data alter the calls in any way. Finally database (4) represents a more real-life case where we do not know in advance what species may be present, so we compare against all available finished and draft bacterial genomes. Note that when we introduce the whole range of genomic data for bacteria, it may be the case that some sequence reads that previously discriminated *S. aureus* from *S. epidermis* (when only genome of those 2 species were in the database) now no longer have that effect because they may map to regions shared by other species that are now in the full comparison database.

An important byproduct will be a practical determination of the LOD of bacterial pathogens in a mock human clinical sample using current NGS techniques: Can clinically-relevant levels be reliably detected and confidently discriminated by unbiased sequencing?

This pilot effort has been a pothole-magnet for many reasons. The strains used had to be at a low safety level; strains available must be sequenced at high quality; acquisition of study reagents can't require onerous legal agreements; scale of experiment cost needs to be kept reasonable; defining clinical-relevance levels is not trivial, etc. The project was hit hard by 2013 Summer Department of Defense furloughs and the Fall government shutdown. However, all samples have been made and sequencing has been performed.

As of this writing, analysis is still underway. Early impressions include that we probably need more than just RefSeq genomes and that draft genomes help analysis more than hurt it (by providing far wider strain coverage.) It is likely that assay and instrument developers will want as much genomic strain coverage as possible (e.g., including draft genomes of some threshold quality based primarily on coverage depth) to perform *in silico* comparisons and to check against lab test results for initial verification. Developers might choose to acquire greater strain coverage from under-represented circulating strains from various geographical regions, to ensure proper assay performance. Depositing the genomes from such strains in public databases should be a pre-requisite for novel genomic data to be presented to any regulatory agency for consideration in a validation study.

Regulatory agencies will likely want to focus on high quality genomes from a set of "type strains" for formal *in silico* validation testing. (Note that this should be a proper subset of the broader database used by developers.) Both assay developers and regulators will need to be aware that increased genomic sequencing of strains will from time to time uncover anomalies:

- Contaminated or mis-labeled strains
- Taxonomic assignment errors at the species level
- Outright failures of classical (non-genomic) taxonomy to deal with the spectrum of genomes that exist
- Sequencing and assembly errors in trusted reference genomes

The most interesting result from the pilot study is that current NGS appears to be potentially insufficient to provide confident bacterial species discrimination at low clinical-relevance levels from mock human samples. The reason for this is simple: the sample is swamped by human DNA reads and there is insufficient pathogen genome coverage at realistic clinical levels to be confident of hitting enough discriminating regions to differentiate the species. Increased sequence depth of evolving NGS instruments is one potential solution to this problem. Targeted amplification of discriminating regions of known agents of clinical interest is another approach. We note that this problem will also exist for the use of NGS to detect viruses in blood or other biological products.

Summary

Bacterial genome reference databases are being developed that will be relevant for diverse needs including:

- Vaccine safety
- Blood safety
- Validating molecular assays for human clinical diagnostics, food and product safety, animal health, environmental monitoring, biodefense, etc.

Different user communities may require customization around a common core of high-quality bacterial genomes. The unique metadata requirements/desires of the different user communities may cause problems with efforts trying to reach a single, common standard. Incomplete metadata will always be a fact of life for bacterial genome reference databases.

Draft bacterial genomes will almost certainly play an increasing role in reference databases, due to the high costs of closing and “finishing” sequencing. Bacterial reference databases for now will have to incorporate much legacy information for which inadequate metadata will be available. However, continuing improvements in sequencing and assembly technologies may prompt extensive future re-sequencing of important legacy bacterial strains, if the need for genome and metadata quality is strong enough to warrant the cost of doing so.